Challenges for enabling Grid Computing over Optical Networks

Cees de Laat

SURFnet

EU

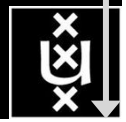
BSIK

NWO

University of Amsterdam

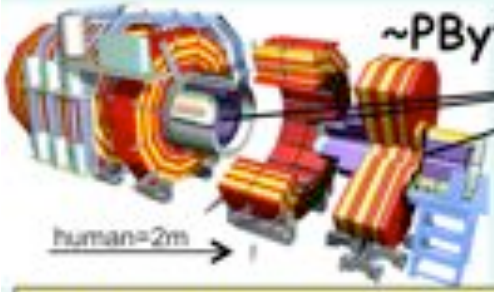
TNO

NCF



LHC Data Grid Hierarchy

CMS as example, Atlas is similar



~PByte/sec

Online System

~100 MBytes/sec

100000 flops/byte

event simulation

Tier 0 +1

HPSS

event reconstruction

CMS detector: 15m X 15m X 22m

12,500 tons, $700M.

~2.5 Gbits/sec

Tier 1

Italian Regional Center

German Regional Center

NIKHEF Dutch Regional Center

FermiLab, USA Regional Center

...

~0.6-2.5 Gbps

analysis

Tier2 Center

Tier 2

~0.6-2.5 Gbps

Tier 3

Institute ~0.25TIPS

Physics data cache

100 - 1000 Mbits/sec

Tier 4

Workstations

Courtesy Harvey Newman, CalTech and CERN

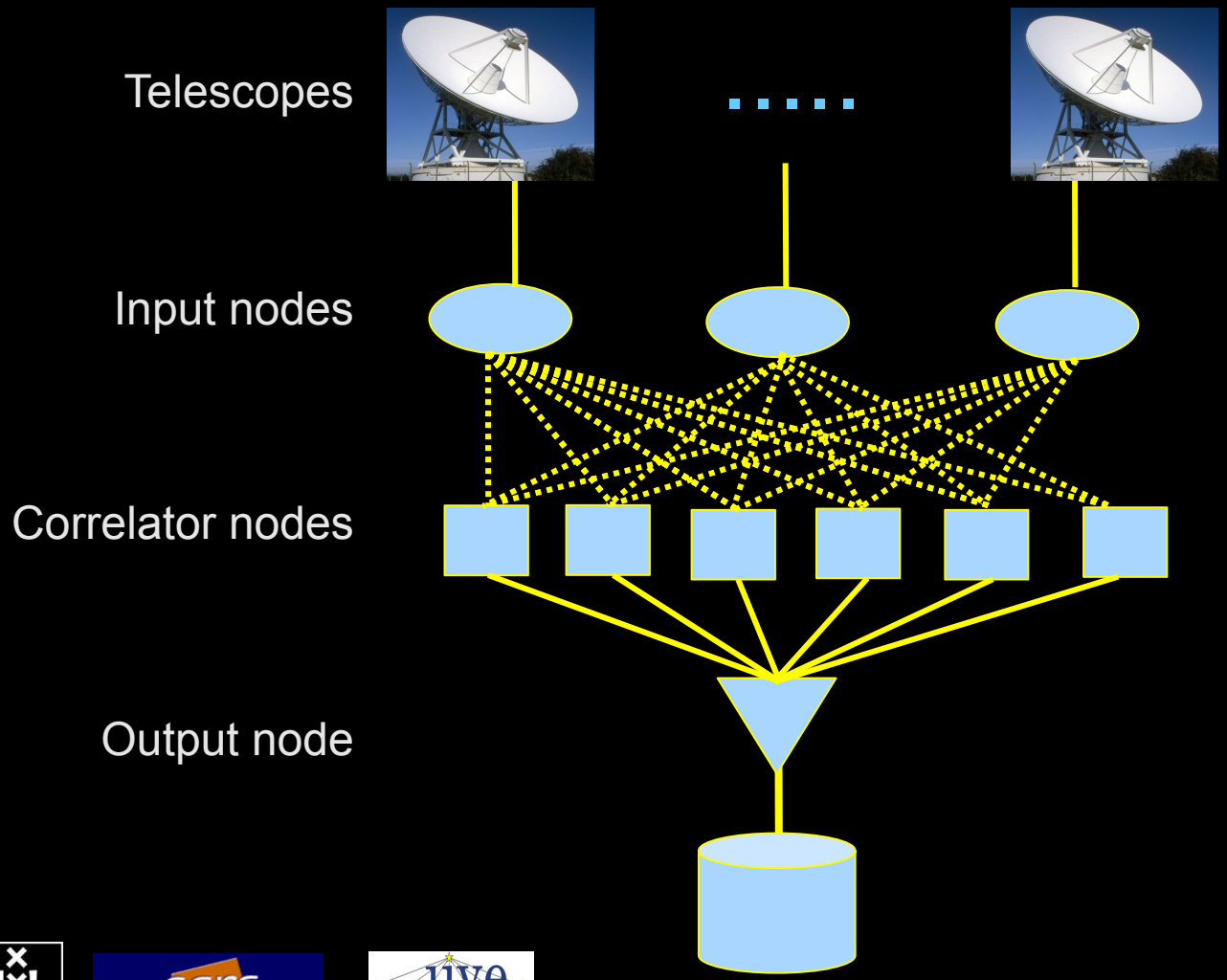
CERN/CMS data goes to 6-8 Tier 1 regional centers, and from each of these to 6-10 Tier 2 centers.

Physicists work on analysis "channels" at 135 institutes. Each institute has ~10 physicists working on one or more channels.

2000 physicists in 31 countries are involved in this 20-year experiment in which DOE is a major player.

The SCARIe project

SCARIe: a research project to create a Software Correlator for e-VLBI.
VLBI Correlation: signal processing technique to get high precision image from spatially distributed radio-telescope.



To equal the hardware correlator we need:

16 streams of 1Gbps

16 * 1Gbps of data

2 Tflops CPU power

2 TFlop / 16 Gbps =

1000 flops/byte

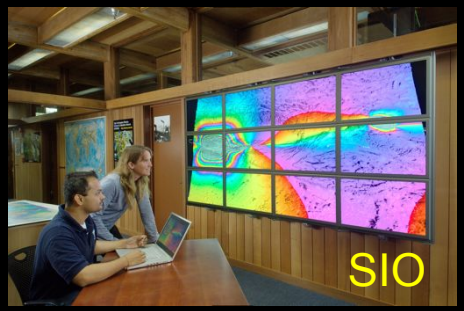
THIS IS A DATA FLOW PROBLEM !!!

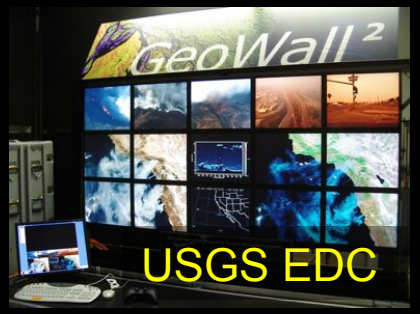

LOFAR as a Sensor Network

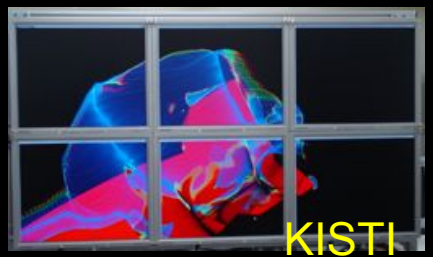
20 flops/byte

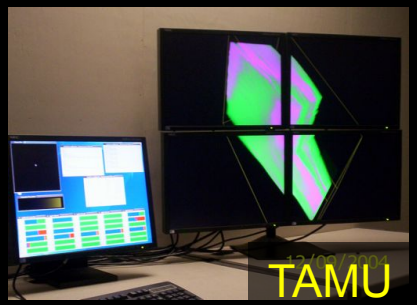
- LOFAR is a large distributed research infrastructure:
  - Astronomy:
    - >100 phased array stations
    - Combined in aperture synthesis array
    - 13,000 small “LF” antennas
    - 13,000 small “HF” tiles
  - Geophysics:
    - 18 vibration sensors per station
    - Infrasound detector per station
  - >20 Tbit/s generated digitally
  - >40 Tflop/s supercomputer
  - innovative software systems
    - new calibration approaches
    - full distributed control
    - VO and Grid integration
    - datamining and visualisation

US and International OptIPortal Sites









Real time, multiple 10 Gb/s

The “Dead Cat” demo

1 Mflops/byte

SC2004,
Pittsburgh,
Nov. 6 to 12, 2004
iGrid2005,
San Diego,
sept. 2005

Many thanks to:
AMC
SARA
GigaPort
UvA/AIR
Silicon Graphics, Inc.
Zoölogisch Museum

M. Scarpa, R.G. Belleman, P.M.A. Sloot and C.T.A.M. de Laat, "Highly Interactive Distributed Visualization", iGrid2005 special issue, Future Generation Computer Systems, volume 22 issue 8, pp. 896-900 (2006).

IJKDIJK

300000 * 60 kb/s * 2 sensors (microphones) to cover all Dutch dikes

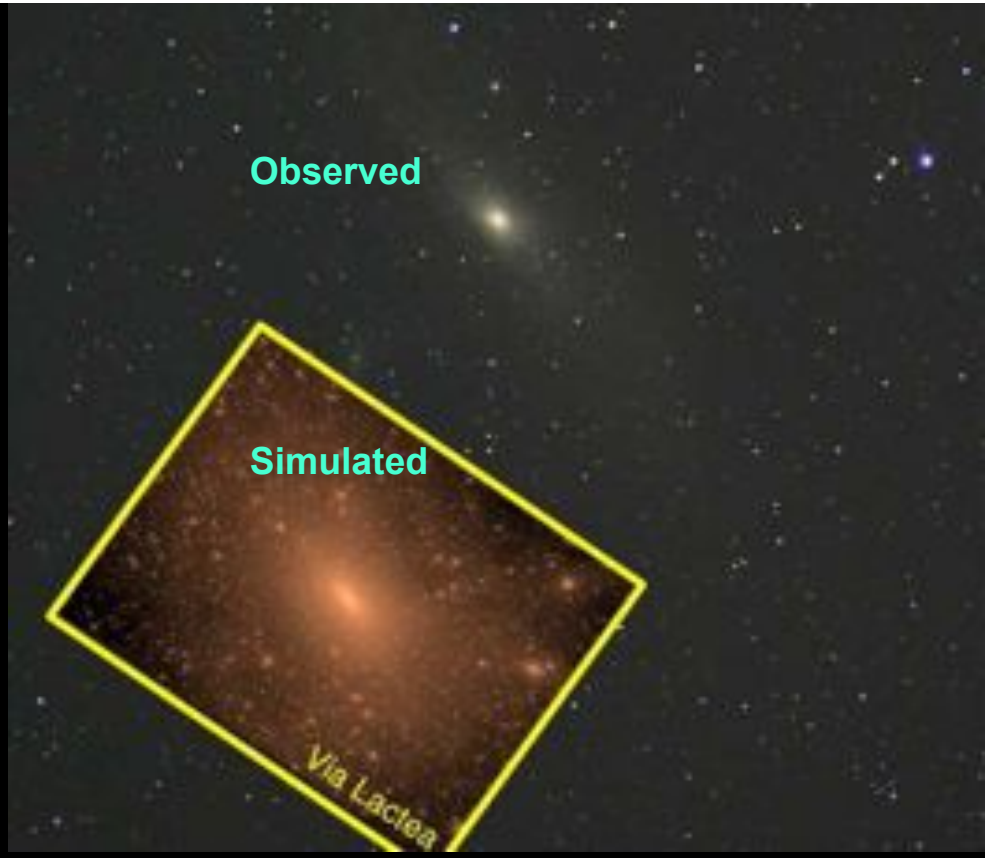
Sensor grid: instrument the dikes

First controlled breach occurred on sept 27th ‘08:

CosmoGrid

- Motivation:

  previous simulations found >100 times more substructure than is observed!



- Simulate large structure formation in the Universe
  - Dark Energy (cosmological constant)
  - Dark Matter (particles)
- Method: Cosmological *N*-body code
- Computation: Intercontinental SuperComputer Grid

The hardware setup

10 Mflops/byte

- 2 supercomputers :
  - 1 in Amsterdam (60Tflops Power6 @ SARA)
  - 1 in Tokyo (30Tflops Cray XD0-4 @ CFCA)
- Both computers are connected via an intercontinental optical 10 Gbit/s network


7.6 Gb/s
CineGrid
CineGrid @ Holland Festival 2007

Why is more resolution is better?

1. More Resolution Allows Closer Viewing of Larger Image
2. Closer Viewing of Larger Image Increases Viewing Angle
3. Increased Viewing Angle Produces Stronger Emotional Response

HDTV (2K)

1920

1080

30°

3.0 × Picture Height

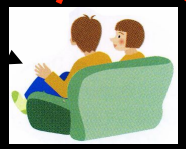
UHDTV(8K)

24 Gb/s

7680

4320

0.75 × Picture Height



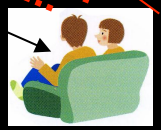
100°

SHD (4K)

3840

2160

7.6 Gb/s



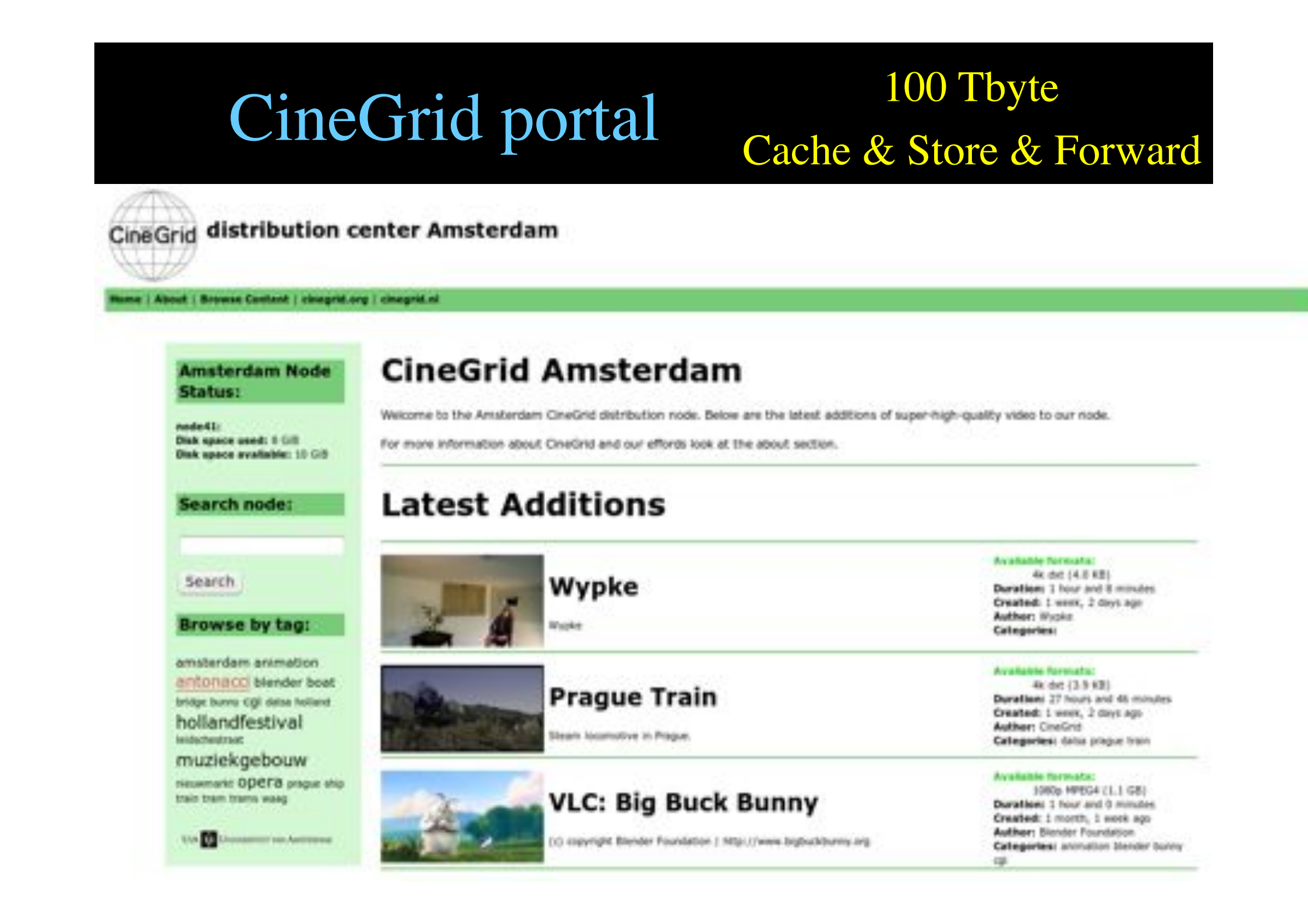
60°

1.5 × Picture Height

Yutaka TANAKA
SHARP CORPORATION
Advanced Image Research Laboratories

CineGrid portal

100 Tbyte
Cache & Store & Forward



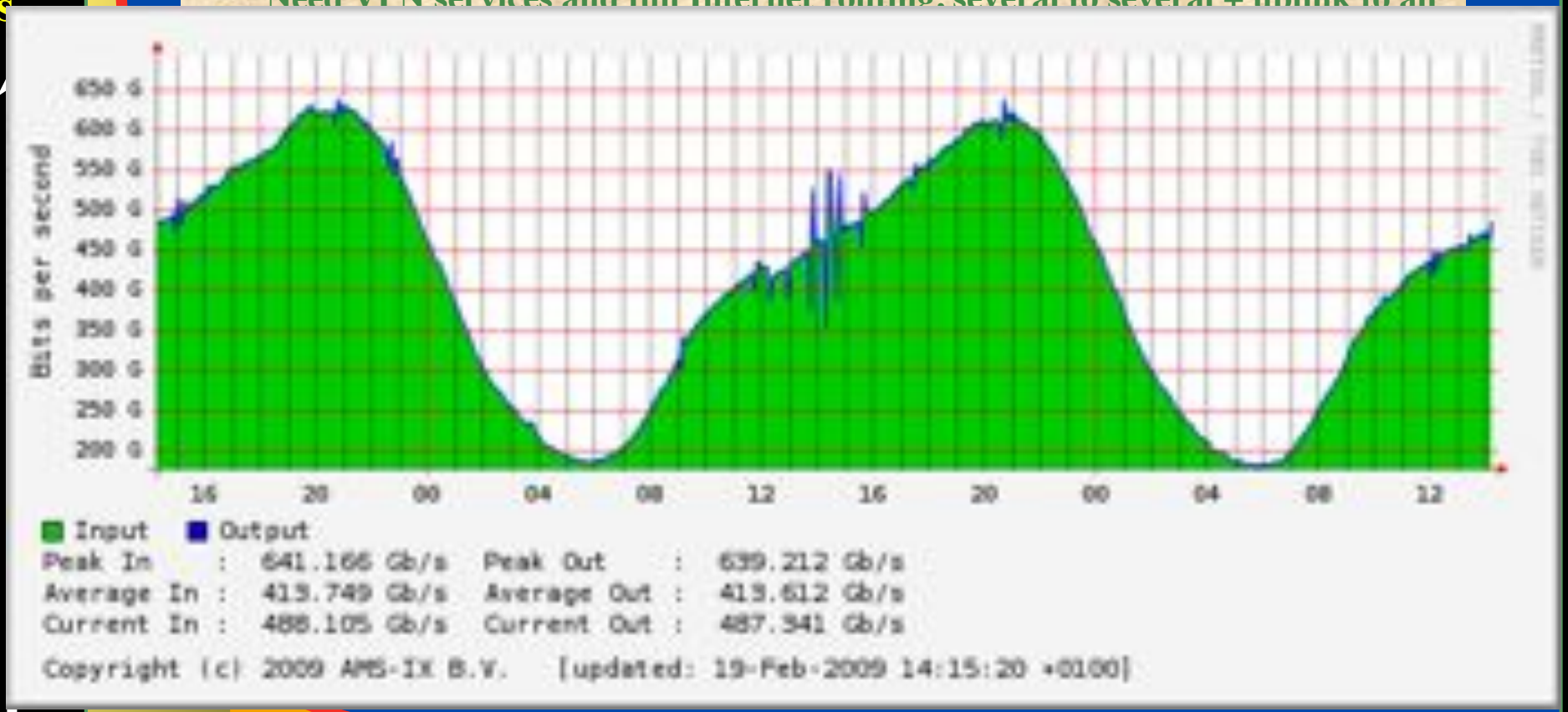
#users

A. Lightweight users, browsing, mailing, home use

Need full Internet routing, one to all

B. Business/grid applications, multicast, streaming, VO's, mostly LAN

Need VPN services and full Internet routing, several to several + uplink to all



B

C

ADSL (12 Mbit/s)

BW

GigE

Ref: Cees de Laat, Erik Radius, Steven Wallace, "The Rationale of the Current Optical Networking Initiatives" iGrid2002 special issue, Future Generation Computer Systems, volume 19 issue 6 (2003)

Towards Hybrid Networking!

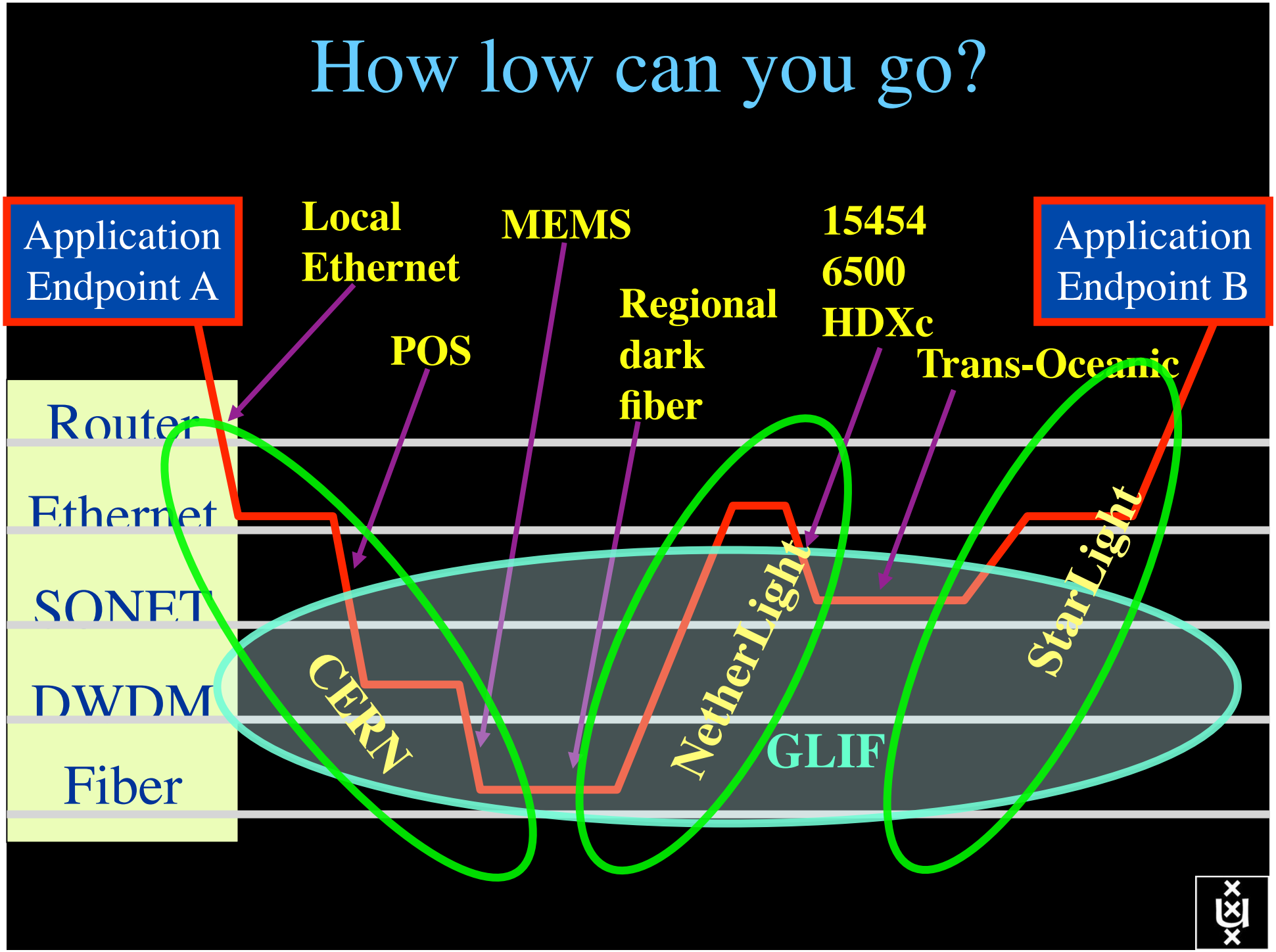
- Costs of photonic equipment 10% of switching 10 % of full routing
  - for same throughput!
  - Photonic vs Optical (optical used for SONET, etc, 10-50 k$/port)
  - DWDM lasers for long reach expensive, 10-50 k$
- Bottom line: look for a hybrid architecture which serves all classes in a cost effective way
  - map A -> L3 , B -> L2 , C -> L1 and L2
- Give each packet in the network the service it needs, but no more !

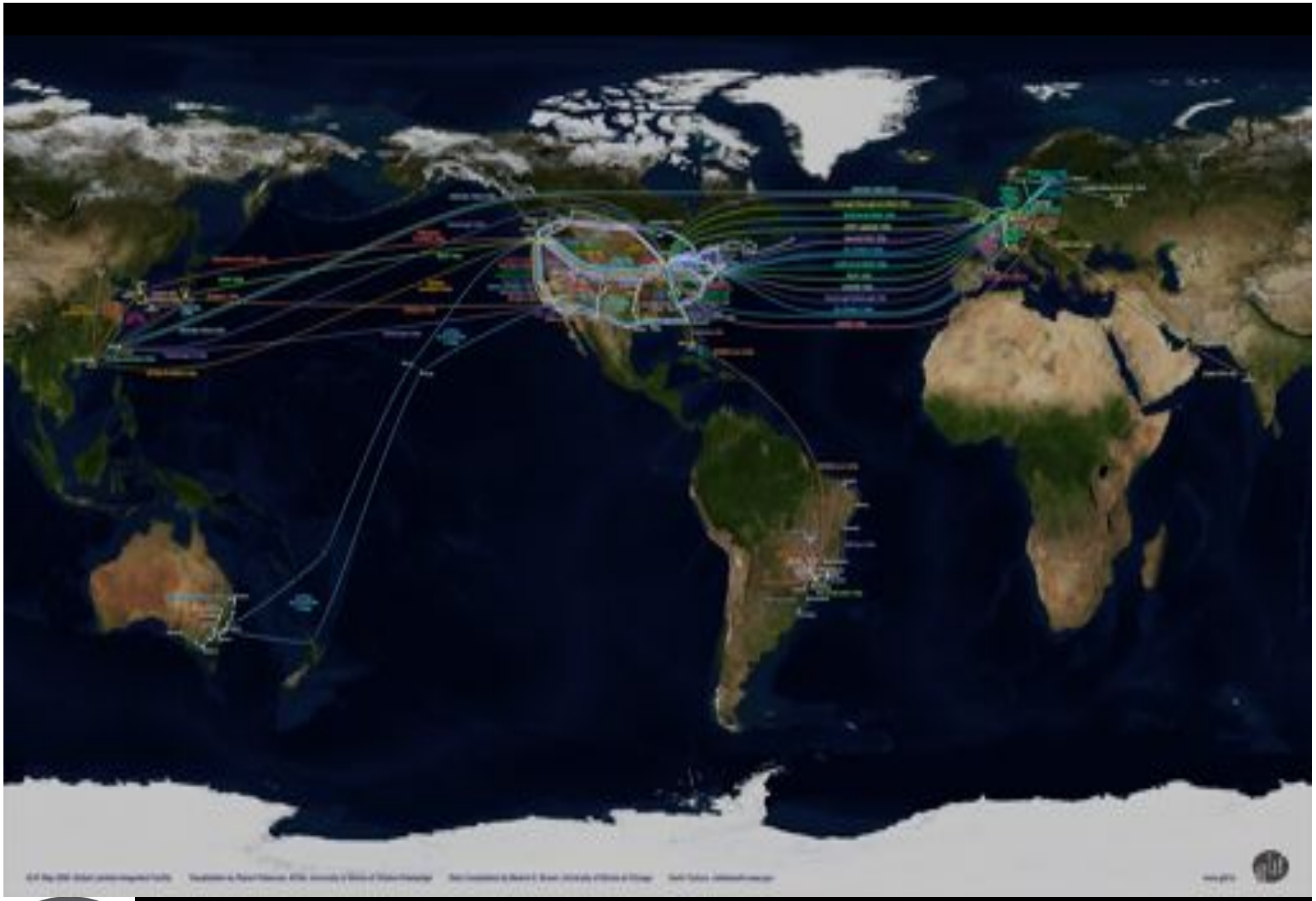
L1 ≈ 2-3 k$/port

L2 ≈ 5-8 k$/port

L3 ≈ 75+ k$/port

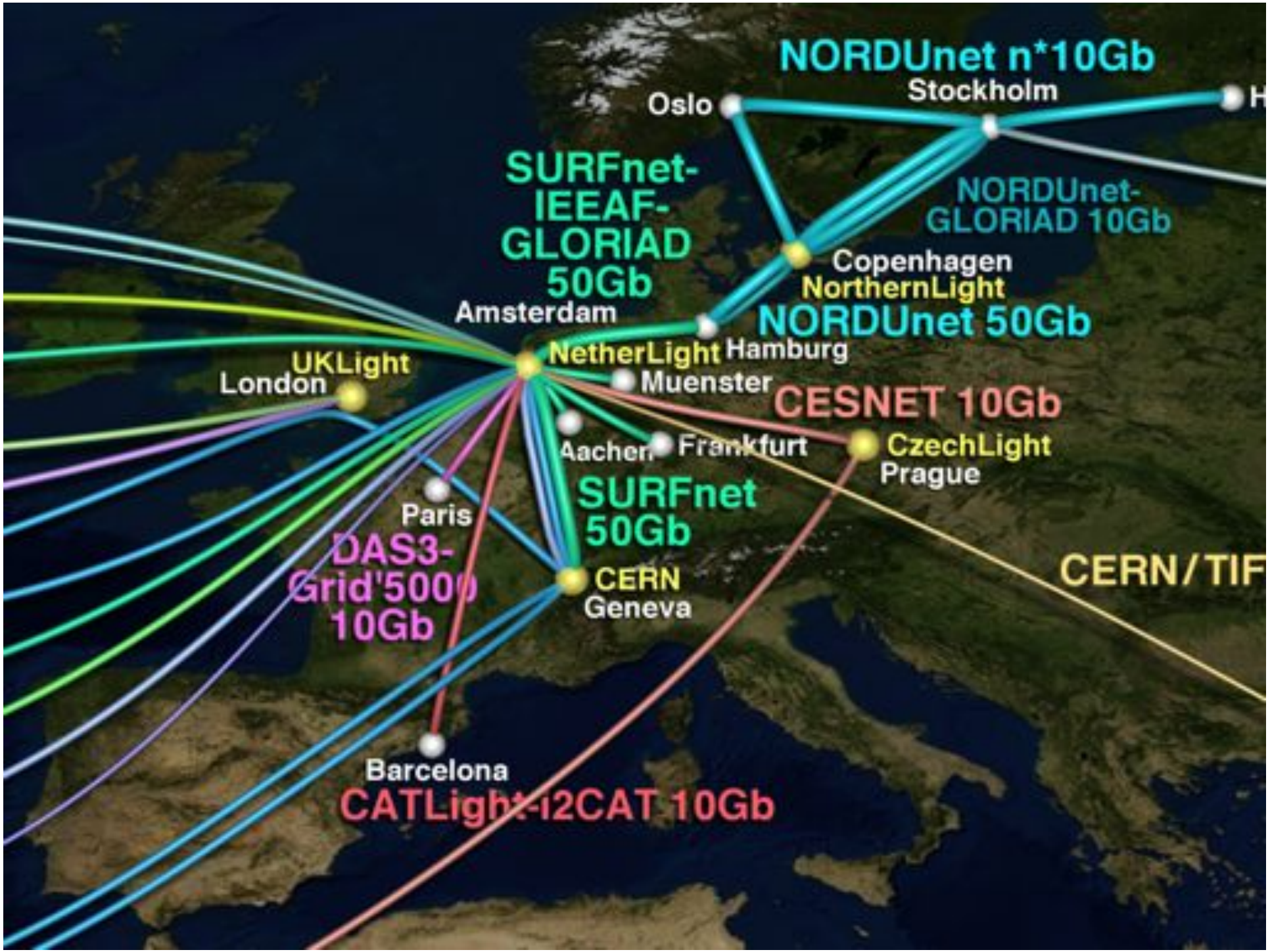
How low can you go?

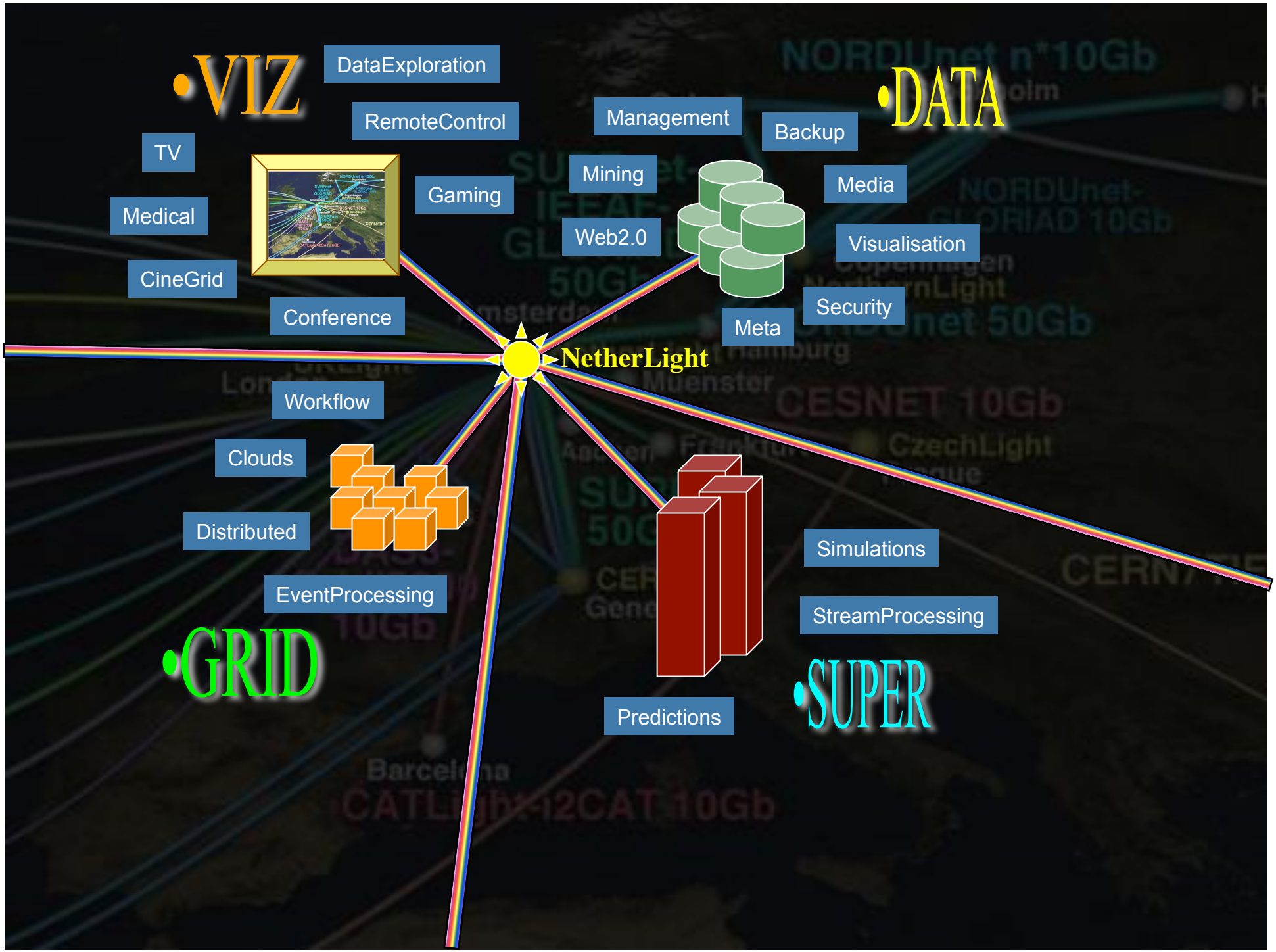




GLIF 2008

Visualization courtesy of Bob Patterson, NCSA
Data collection by Maxine Brown.


NORDUnet n*10Gb
Stockholm
Oslo
SURFnet-
IEEAF-
GLORIAD
50Gb
NORDUnet-
GLORIAD 10Gb
Copenhagen
NorthernLight
NORDUnet 50Gb
Amsterdam
NetherLight
Hamburg
UKLight
London
Muenster
CESNET 10Gb
Aachen
Frankfurt
CzechLight
Prague
SURFnet
50Gb
Paris
DAS3-
Grid'5000
10Gb
CERN
Geneva
CERN/TIF
Barcelona
CATLight-i2CAT 10Gb


•VIZ
DataExploration
RemoteControl
TV
Medical
CineGrid
Conference
Gaming
•DATA
Management
Backup
Mining
Media
Web2.0
Visualisation
Security
Meta
NetherLight
Workflow
Clouds
Distributed
EventProcessing
•GRID
Simulations
StreamProcessing
Predictions
•SUPER

SURFnet

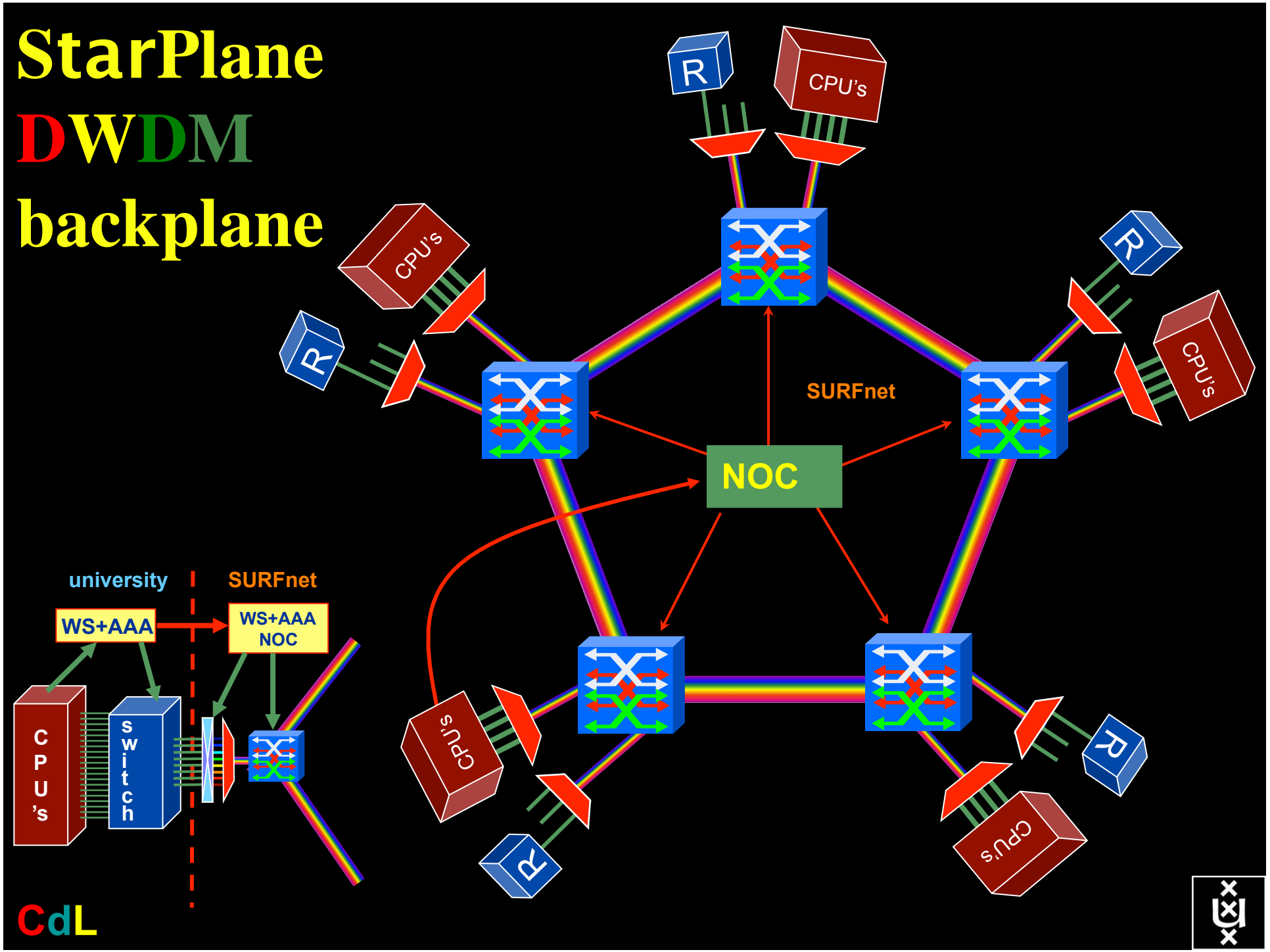
In The Netherlands SURFnet connects between 180:

- universities;
- academic hospitals;
- most polytechnics;
- research centers.

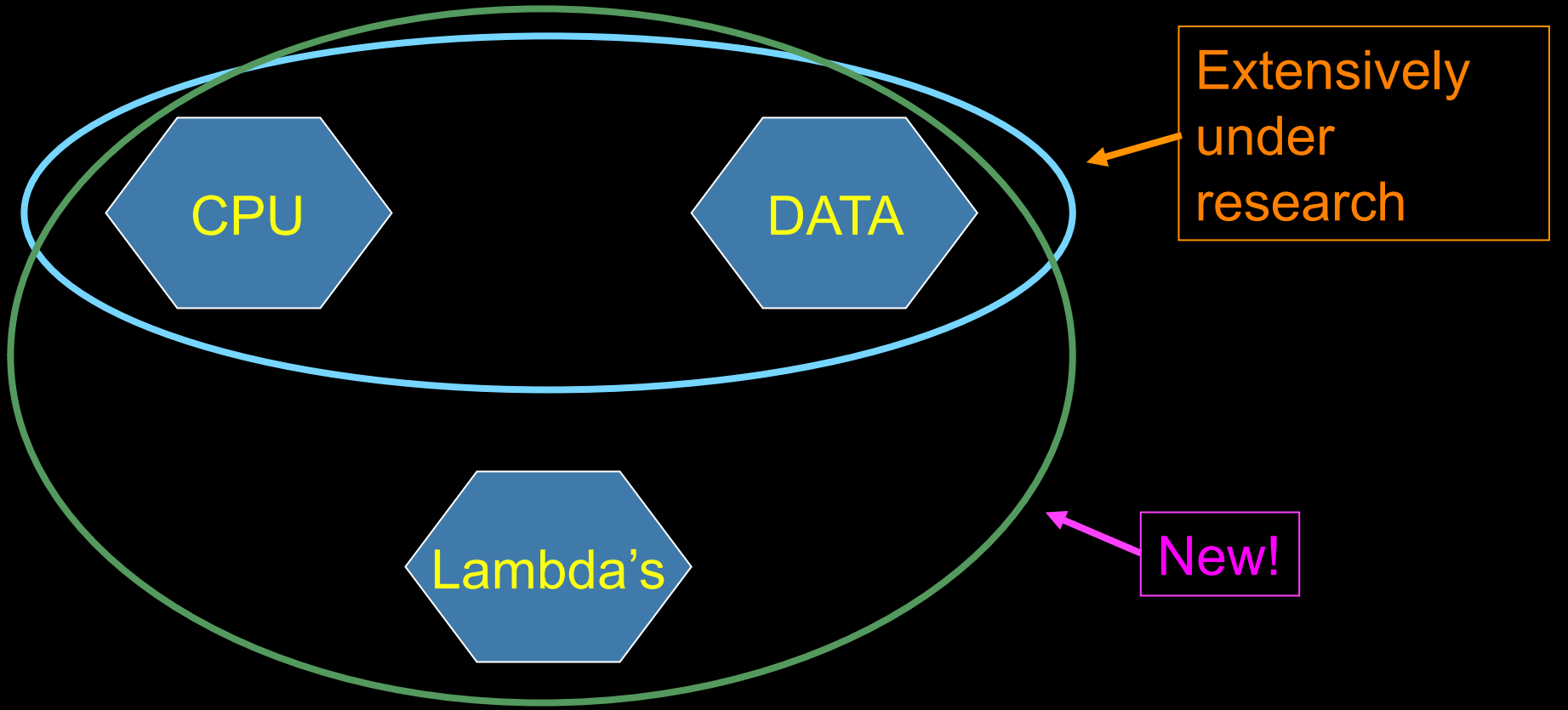
with an indirect ~750K user base

~ 8860 km

scale comparable to railway system


StarPlane
DWDM
backplane
R
CPU's
SURFnet
NOC
university
SURFnet
WS+AAA
WS+AAA
NOC
CPU's
switch
CdL

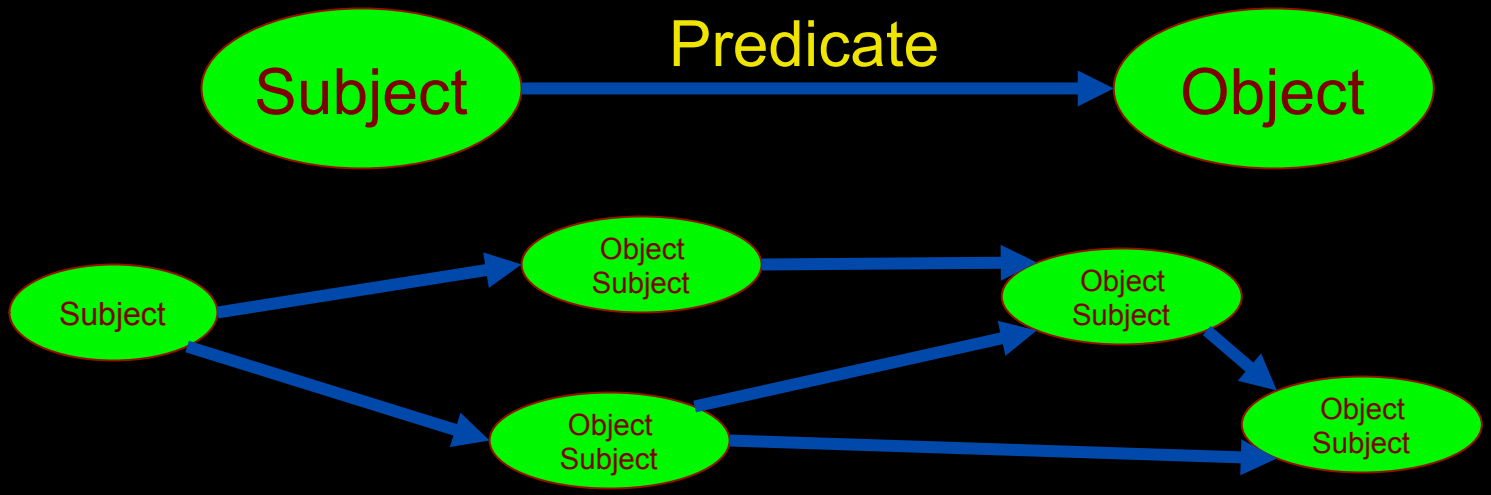
GRID Co-scheduling problem space

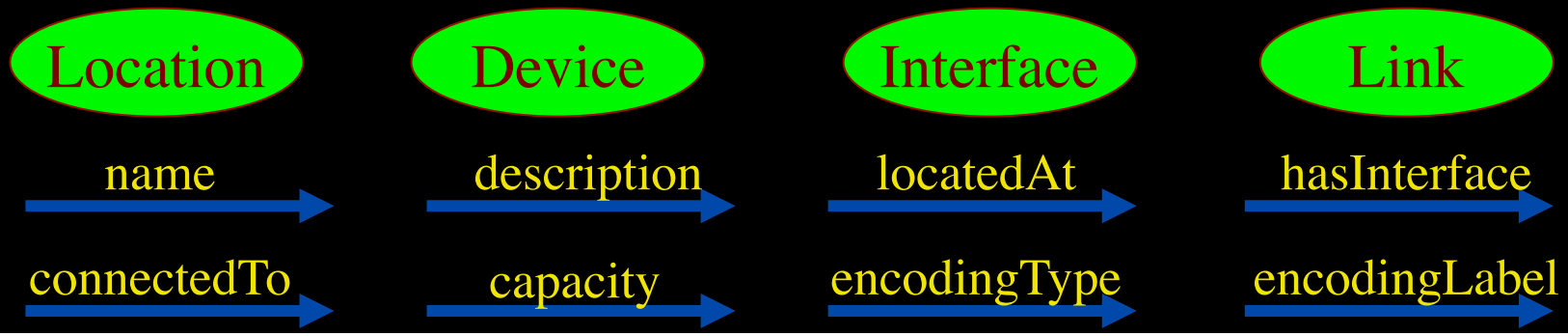


The StarPlane vision is to give flexibility directly to the applications by allowing them to choose the logical topology in real time, ultimately with sub-second lambda switching times on part of the SURFnet6 infrastructure.

Network Description Language

- From semantic Web / Resource Description Framework.
- The RDF uses XML as an interchange syntax.
- Data is described by triplets:



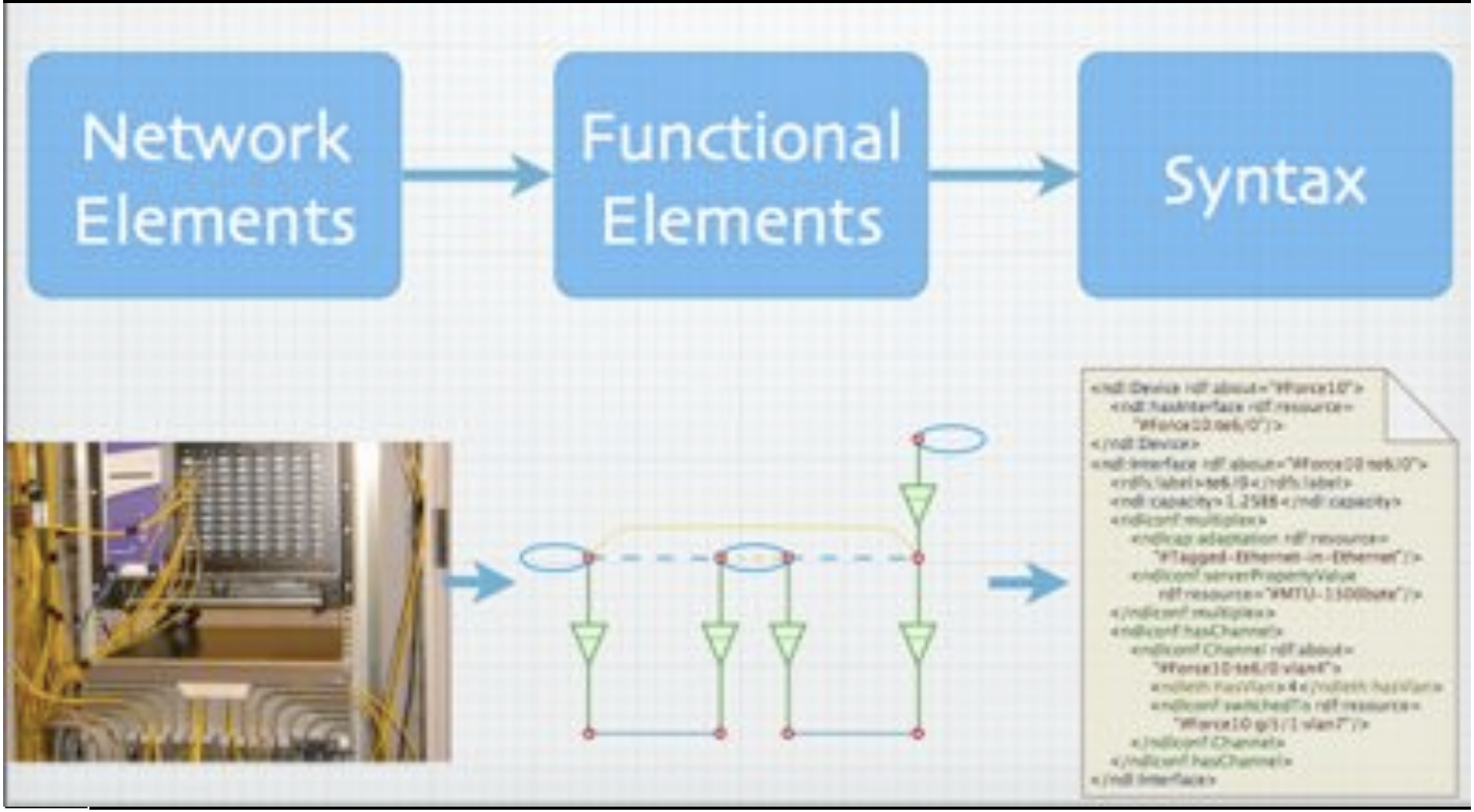


Network Description Language

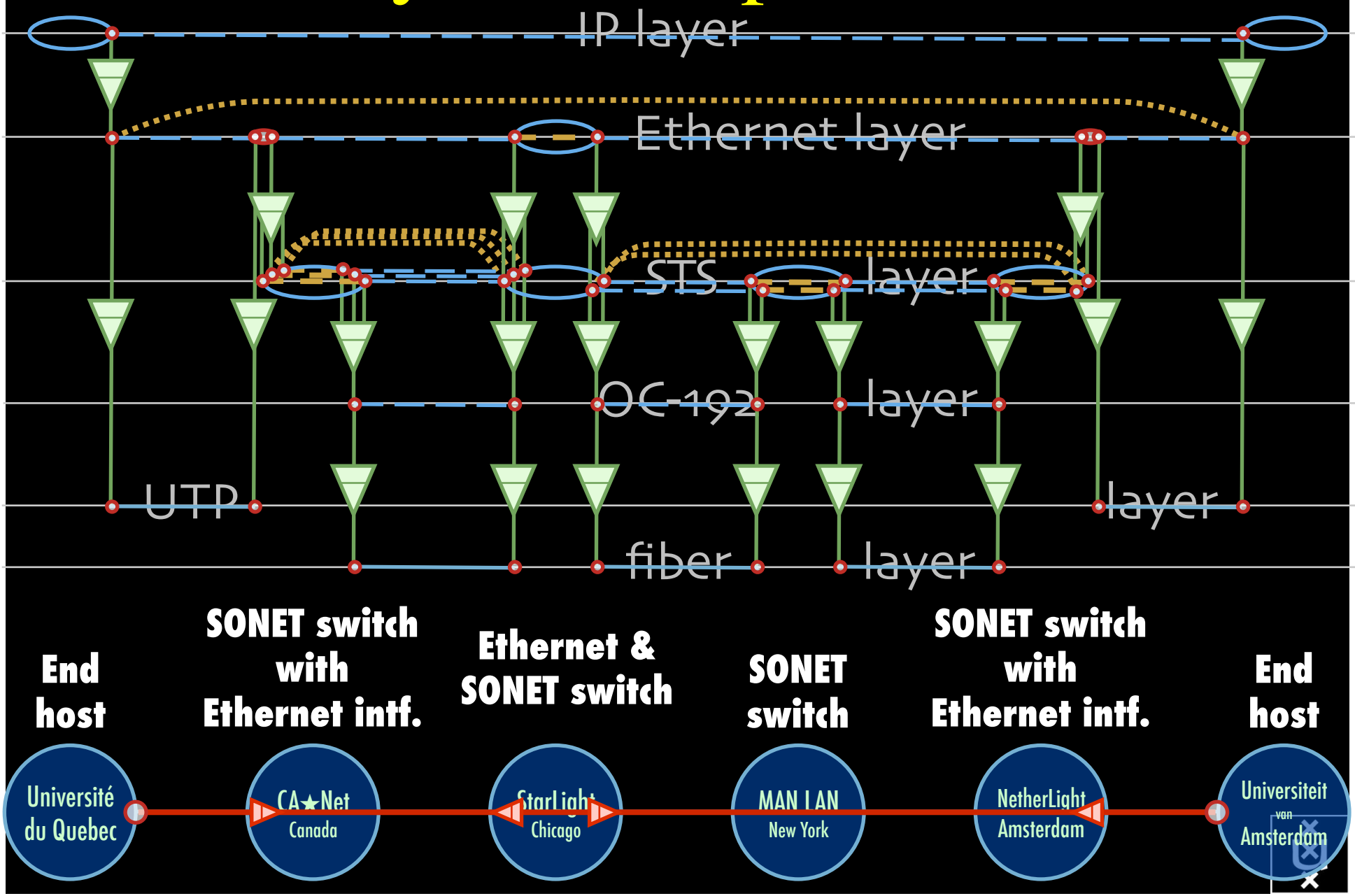
Choice of RDF instead of flat XML descriptions

Grounded modeling based on G0805 description:

Article: F. Dijkstra, B. Andree, K. Koymans, J. van der Ham, P. Grosso, C. de Laat, *"A Multi-Layer Network Model Based on ITU-T G.805"*



Multi-layer descriptions in NDL

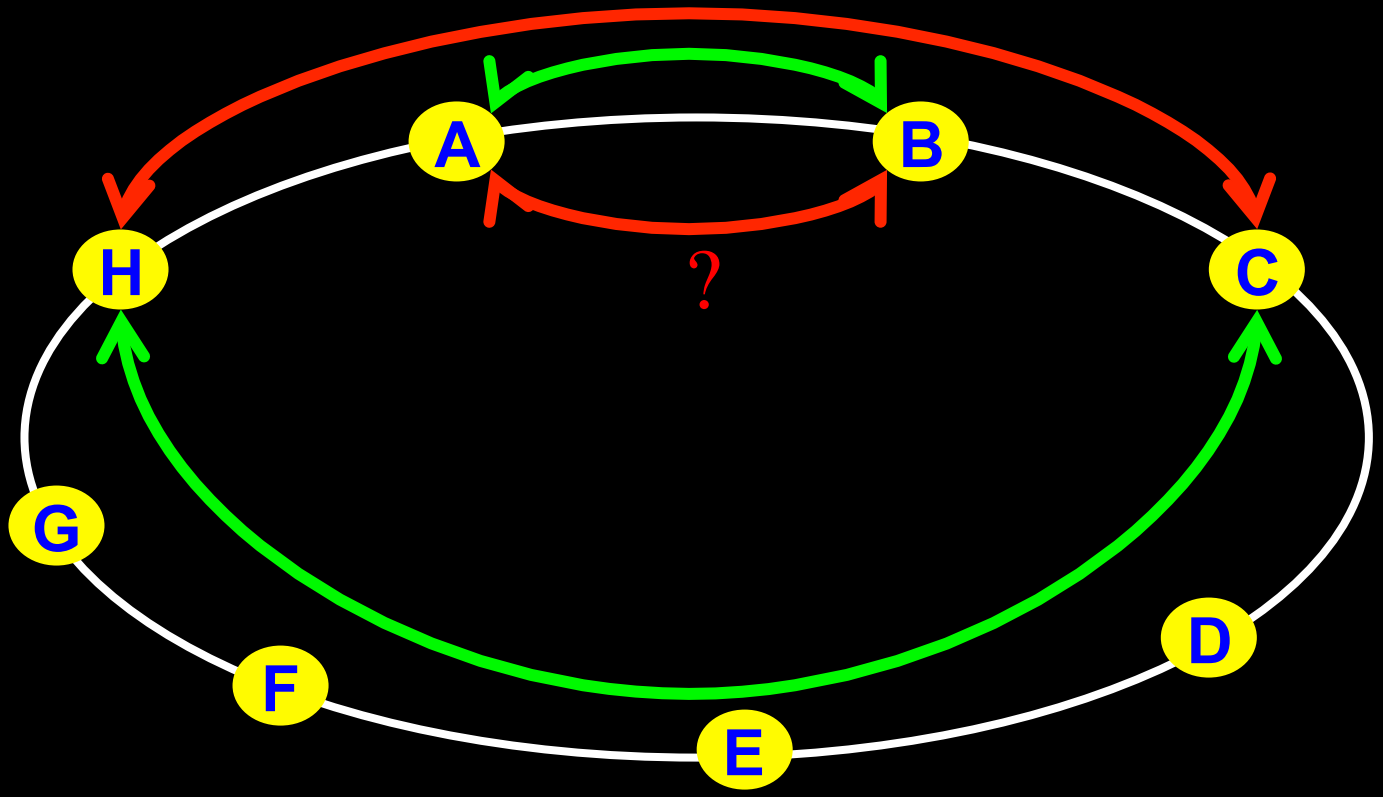


The Problem

I want HC and AB

Success depends on the order

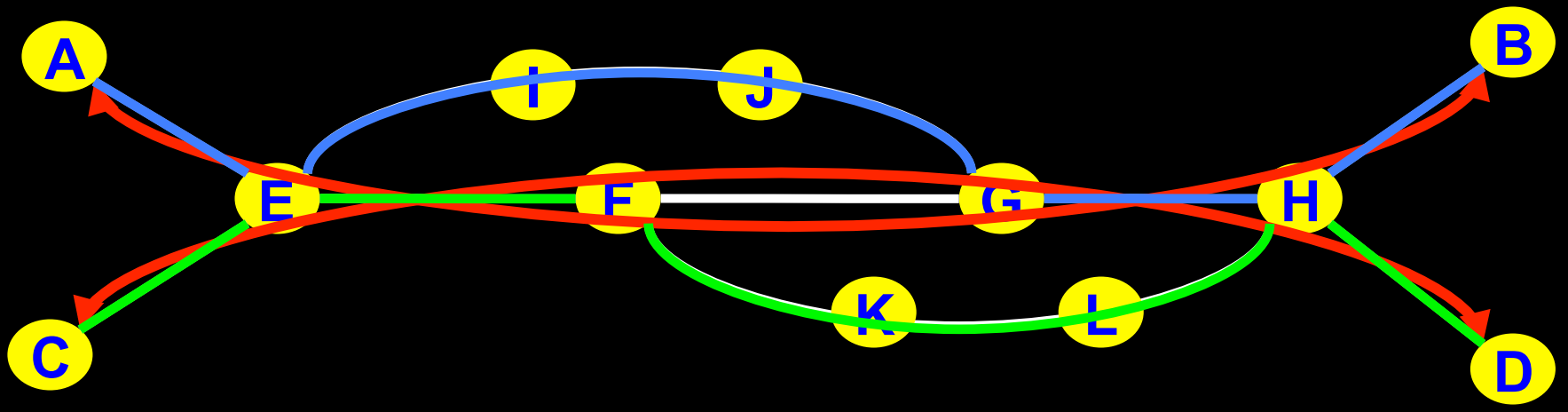
Wouldn't it be nice if I could request [HC, AB, ...]



Another one ☺

I want AB and CD

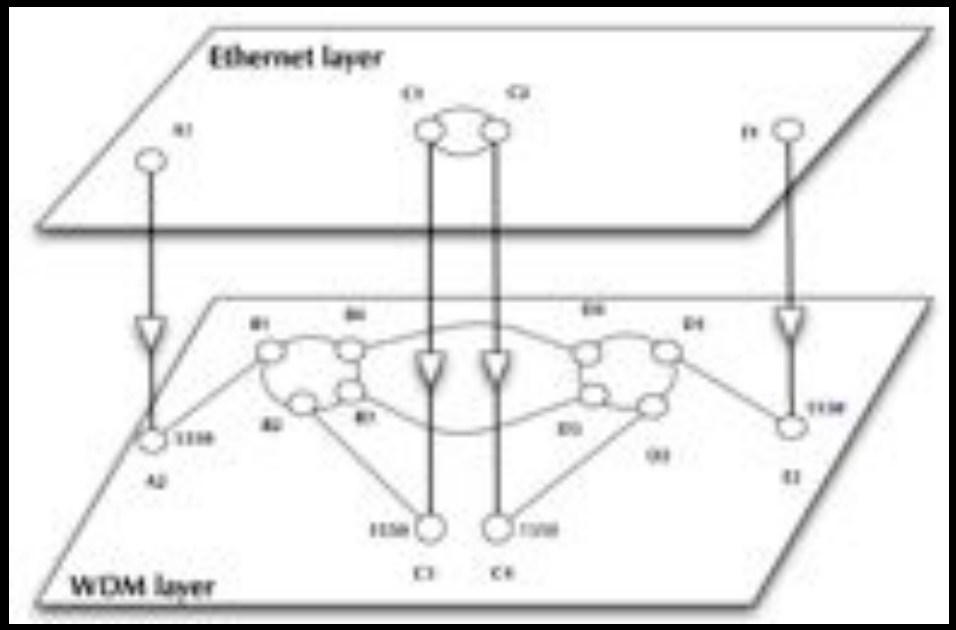
Success does not even depend on the order!!!

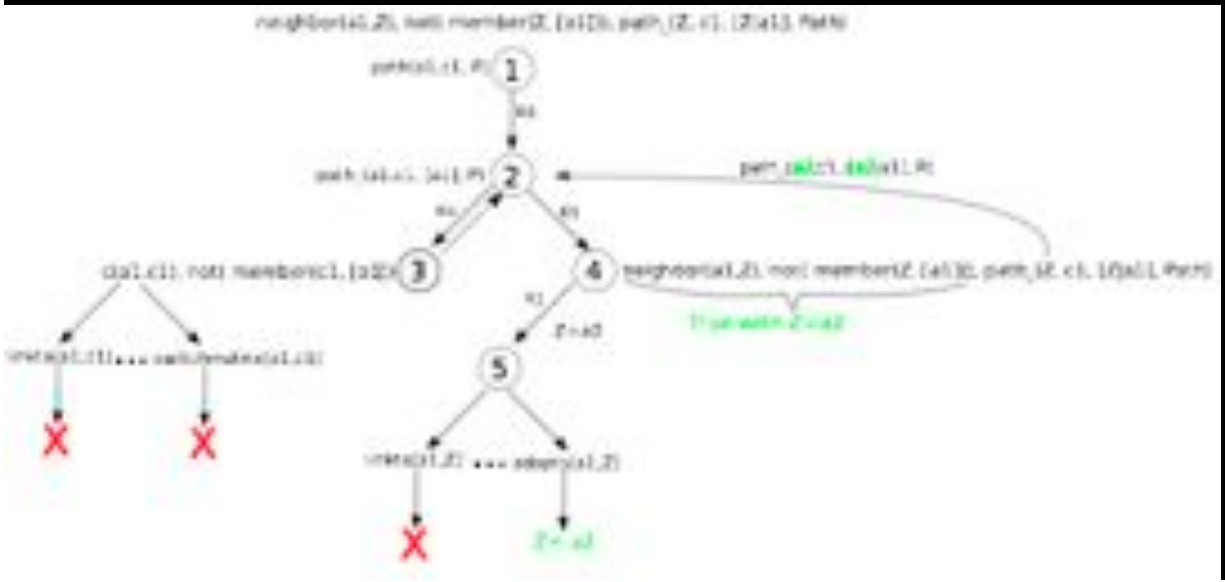


NDL + PROLOG

Research Questions:

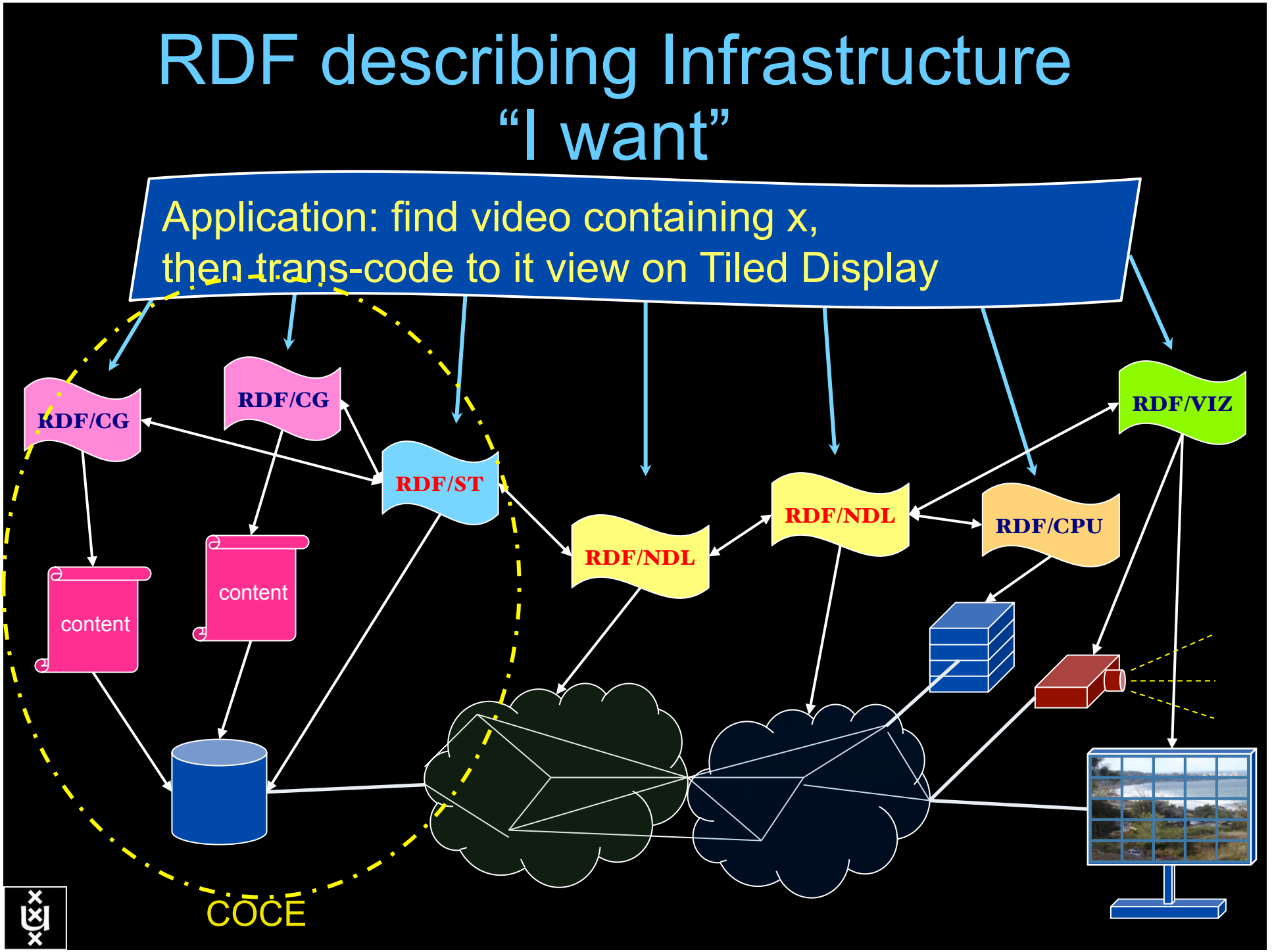
- order of requests
- complex requests
- usable leftovers





- Reason about graphs
- Find sub-graphs that comply with rules

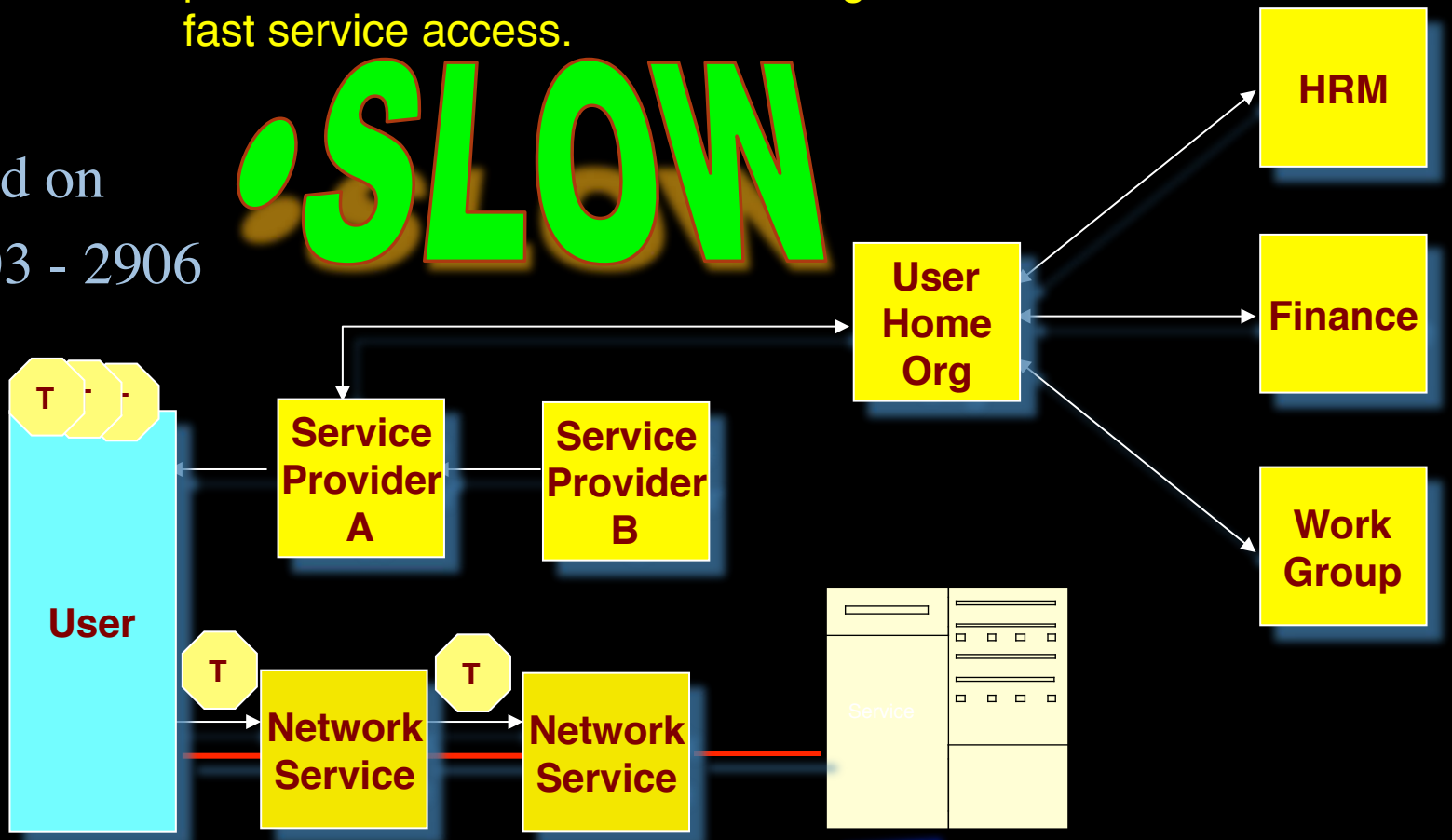
RDF describing Infrastructure
"I want"



Use AAA concept to split (time consuming) service authorization process from service access using secure tokens in order to allow fast service access.

Based on
RFC 2903 - 2906

- SLOW



- fast

ref Leon Gommans, Li Xu, Fred Wan, Yuri Demchenko, Mihai Cristea, Robert Meijer, Cees de Laat ,
"Multi-Domain Lightpath Authorization using Tokens", FGCS, Vol 25, issue 2, feb 2009, pages 153-160.

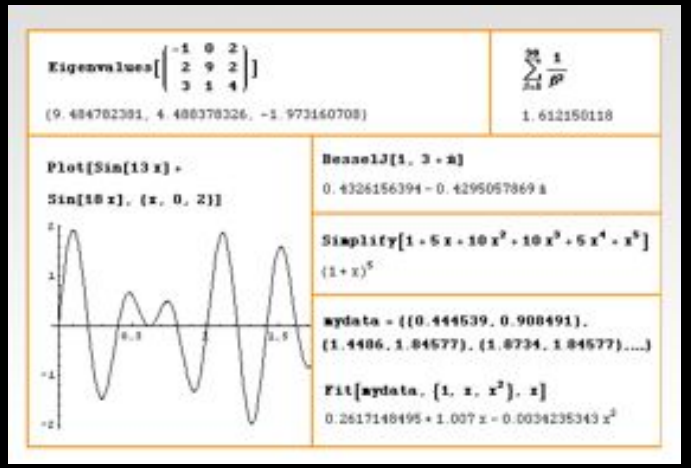
TeraThinking

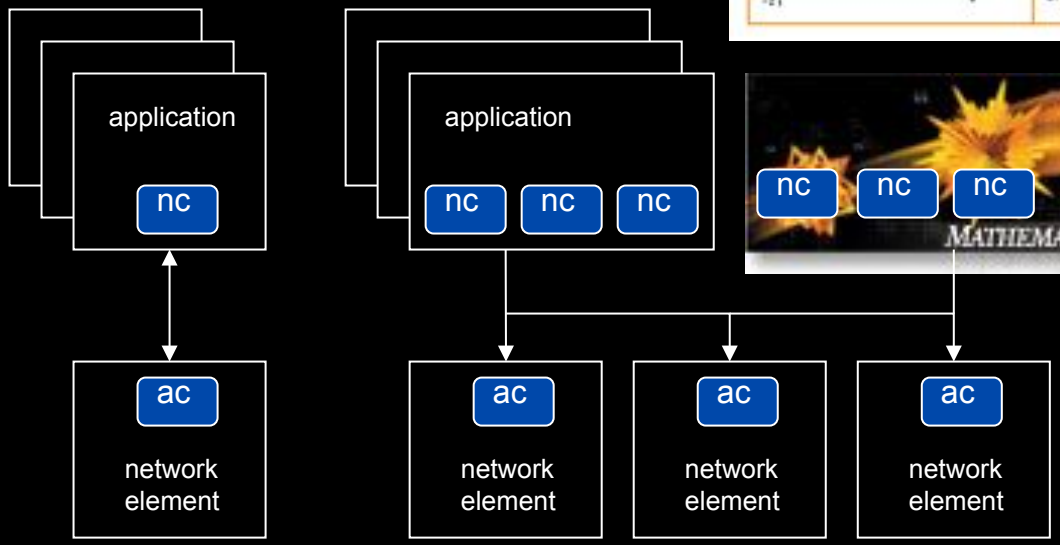
- What constitutes a Tb/s network?
- CALIT2 has 8000 Gigabit drops ?->? Terabit Lan?
- look at 80 core Intel processor
  - cut it in two, left and right communicate 8 TB/s
- think back to teraflop computing!
  - MPI turns a room full of pc's in a teraflop machine
- massive parallel channels in hosts, NIC's
- TeraApps programming model supported by
  - TFlops -> MPI / Globus
  - TBytes -> OGSA/DAIS
  - TPixels -> SAGE
  - TSensors -> LOFAR, LHC, LOOKING, CineGrid, ...
  - Tbit/s -> ?

ref Larry Smarr & CdL

User Programmable Virtualized Networks allows the results of decades of computer science to handle the complexities of application specific networking.

- The network is virtualized as a collection of resources
- UPVNs enable network resources to be programmed as part of the application
- Mathematica, a powerful mathematical software system, can interact with real networks using UPVNs





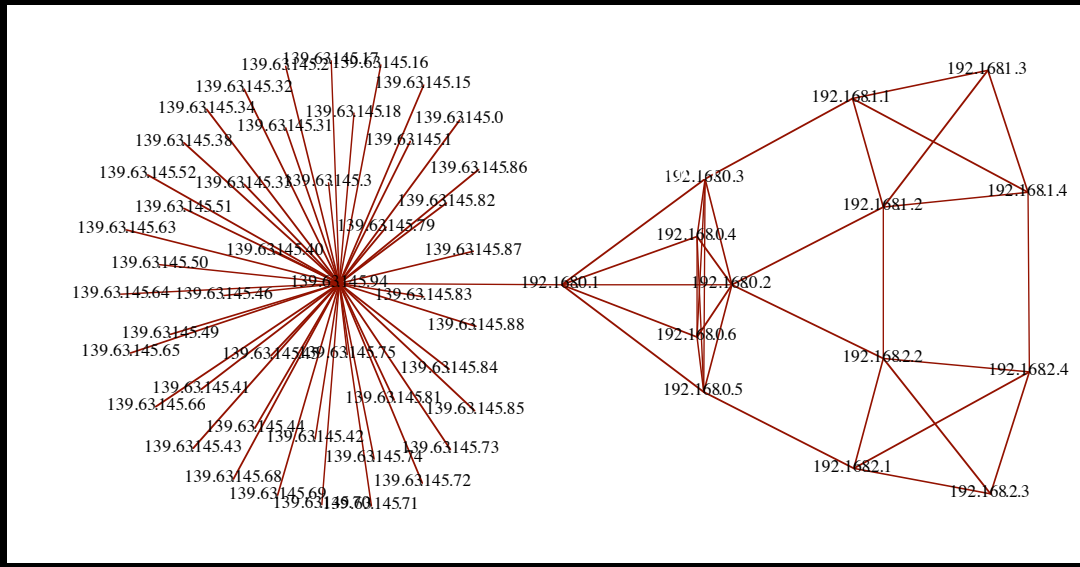
StarPlane

Mathematica enables advanced graph queries, visualizations and real-time network manipulations on UPVNs

Topology matters can be dealt with algorithmically
Results can be persisted using a transaction service built in UPVN

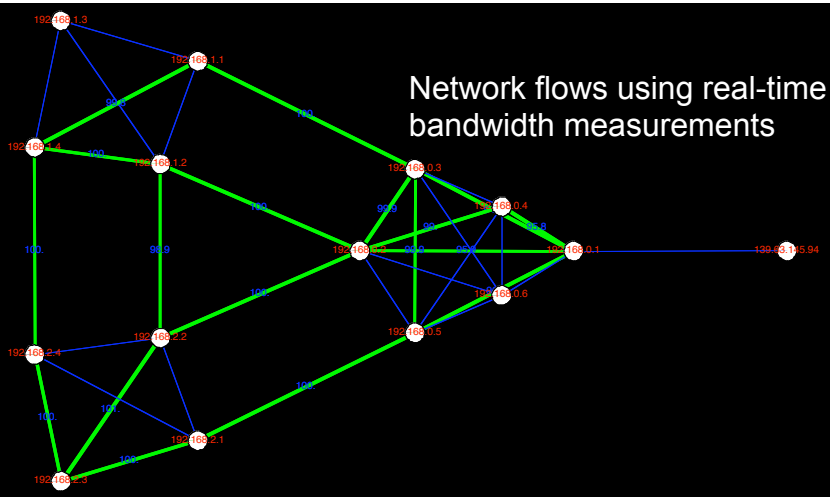
Initialization and BFS discovery of NEs

```
Needs["WebServices`"]
<<DiscreteMath`Combinatorica`
<<DiscreteMath`GraphPlot`
InitNetworkTopologyService["edge.ict.tno.nl"]

Available methods:

{DiscoverNetworkElements,GetLinkBandwidth,GetAllIpLinks,Remote,
NetworkTokenTransaction}

Global`upvnverbose = True;

AbsoluteTiming[nes = BFSDiscover["139.63.145.94"];][[1]]

AbsoluteTiming[result = BFSDiscoverLinks["139.63.145.94", nes];][[1]]

Getting neigbours of: 139.63.145.94
Internal links: {192.168.0.1, 139.63.145.94}
(...)
Getting neigbours of:192.168.2.3

Internal links: {192.168.2.3}
```



Transaction on shortest path with tokens

```
nodePath = ConvertIndicesToNodes[
           ShortestPath[     g,
                     Node2Index[nids,"192.168.3.4"],
                     Node2Index[nids,"139.63.77.49"]],
                     nids];
Print["Path: ", nodePath];
If[NetworkTokenTransaction[nodePath, "green"]==True,
      Print["Committed"], Print["Transaction failed"]];

Path:
{192.168.3.4,192.168.3.1,139.63.77.30,139.63.77.49}

Committed
```



Network flows using real-time bandwidth measurements

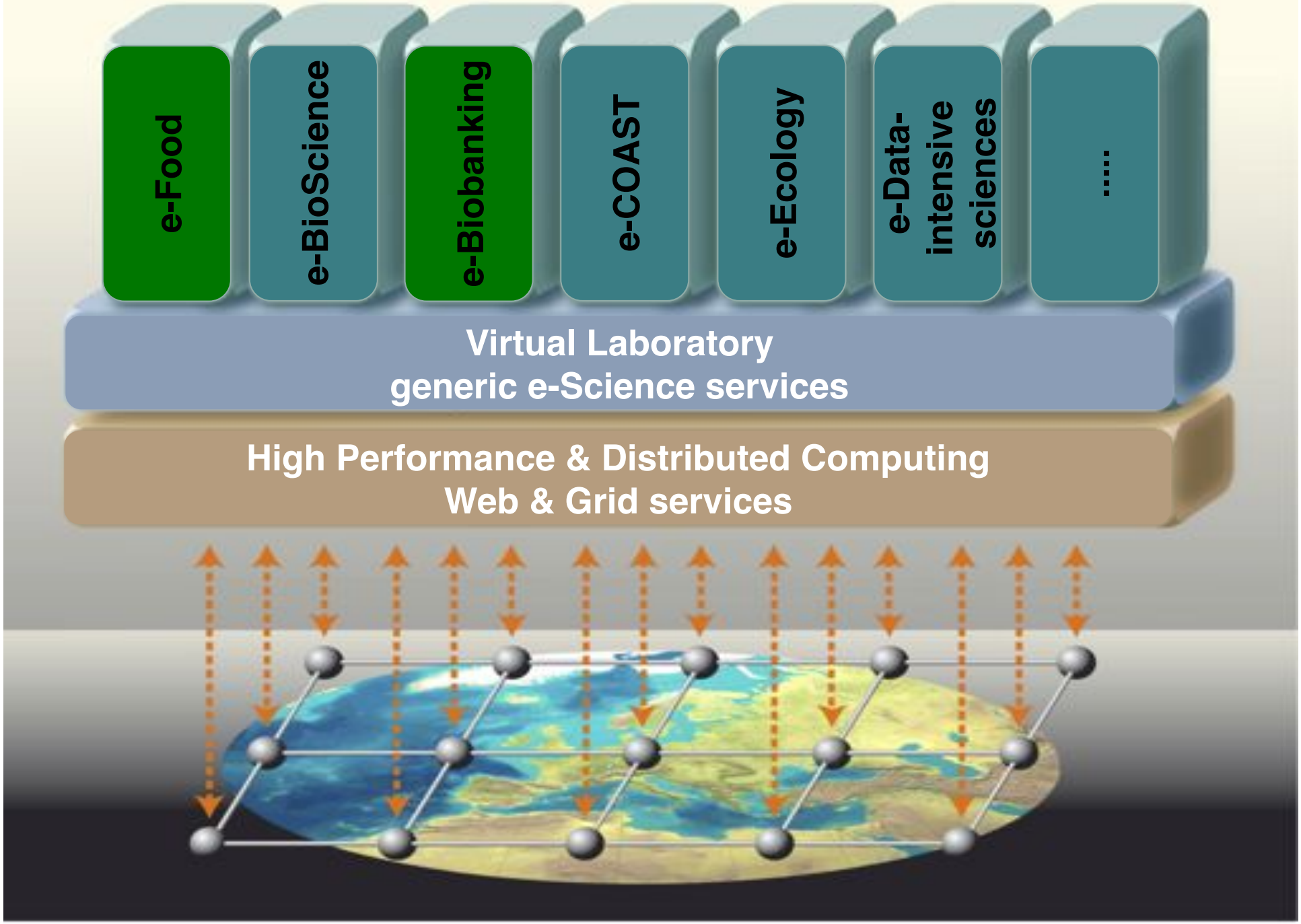
ref: Robert J. Meijer, Rudolf J. Strijkers, Leon Gommans, Cees de Laat, User Programmable Virtualiized Networks, accepted for publication to the IEEE e-Science 2006 conference Amsterdam.

StarPlane

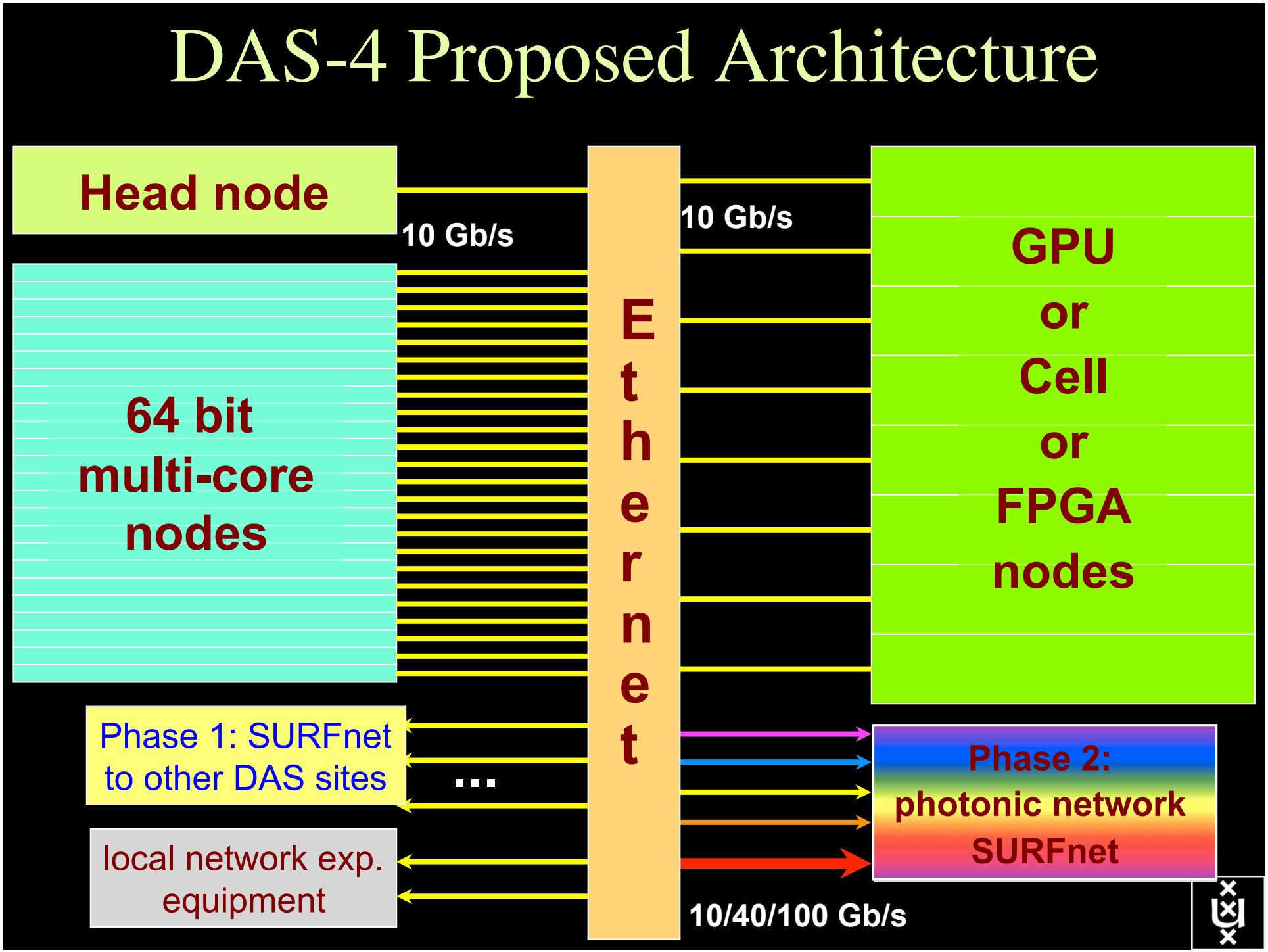
Interactive programmable networks

Themes

- eScience infrastructure virtualization
- Photonic networking -> Tb/s
- Capacity & Capability
- Data handling, integrity, security, privacy
- Cloud paradigm, green compute&store&net&viz
- ENERGY dependency! (2009: 1Wy=1€)


e-Food
e-BioScience
e-Biobanking
e-COAST
e-Ecology
e-Data-intensive sciences
.....
Virtual Laboratory
generic e-Science services
High Performance & Distributed Computing
Web & Grid services

DAS-4 Proposed Architecture



n.a.v. interview met Kees Neggers (SURFnet) & Cees de Laat (UvA)

The COOK Report on Internet Protocol

Technology, Economics, and Policy

ICT and E-Science as an Innovation Platform in The Netherlands

A National Research and Innovation Network What Can the US Learn from Dutch Experience?

"The dogmas of the quiet past are inadequate to the stormy present. As our case is new, so we must think anew and act anew." **Abraham Lincoln**

By means of an examination of research networks in Holland, this issue presents some ideas for ways in which an American National Research, Education and Innovation Network could be structured.

possible are carried out by decentralized groups.

Volume XVII, No. 11
February 2009
ISSN 1871-4331

The Basis for a Future Internet?

Optical Hybrid Networks and e-Science as Platforms for Innovation and Tech Transfer

Editor's Note: I continued the discussion begun on No- slide shows our organization within the University and the search department of KPN. He did a lot of virtualization

- **BSIK projects GigaPort &**
- **VL-e / e-Science**

cookreport.com

Questions ?

A Declarative Approach to Multi-Layer Path Finding Based on Semantic Network Descriptions.

http://delaat.net:/~delaat/papers/declarative_path_finding.pdf

Thanks: Paola Grosso & Jeroen vd Ham & Freek Dijkstra & team for several of the slides.